NAMED ENTITY EXTRACTION FROM SPEECH

Francis Kubala, Richard Schwartz, Rebecca Stone, and Ralph Weischedel

BBN Technologies
70 Fawcett Street
Cambridge, MA 02138

ABSTRACT

We report results using a hidden Markov model to extract information from broadcast news. IdentiFinder™ was trained on the broadcast news corpus and tested on both the 1996 HUB-4 development test data and the 1997 HUB-4 evaluation test data with respect to the named entity (NE) task: extracting

- names of locations, persons, and organizations;
- dates and times;
- monetary amounts and percentages.

Evaluation is based on automatic word alignment of the speech recognition output (the NIST algorithm) followed by the MUC-6/MUC-7 scorer for NE on text, since MUC scoring assumes identical text in the system output and in the answer key. Additionally, we used the experimental MITRE scoring metric (Burger, et al., 1998).

The most encouraging result is that a language-independent, trainable information extraction algorithm degraded on speech input at most by the word error rate of the recognizer.

1. MOTIVATING FACTORS

One of the reasons behind this effort is to go beyond speech transcription (e.g. beyond the dictation problem) to address (at least) shallow understanding of speech. As a result of this effort, we believe that evaluating named entity (NE) extraction from speech offers a measure complementary to word error rate (wer) and represents a measure of understanding. The scores for NE from speech seem to track quality of speech recognition proportionally, i.e., NE performance degrades at worst linearly with word error rate.

A second motivation is the fact that NE is the first information extraction task from text showing success, with error rates on newswire less than 10%. The named entity problem has generated much interest, as evidenced by its inclusion as an understanding task to be evaluated in both the Sixth and Seventh Message Understanding Conferences (MUC-6 and MUC-7), in the First and Second Multilingual Entity Task evaluations (MET-1 and MET-2), and as a planned track in the next broadcast news evaluation. Furthermore, at least one commercial product has emerged: NameTag™ from IsoQuest. NE is defined by a set of annotation guidelines, an evaluation metric, and example data (Chinchor, 1997).

2. THE NAMED ENTITY PROBLEM FOR SPEECH

The named entity task is to identify all named locations, named persons, named organizations, dates, times, monetary amounts, and percentages. Though this sounds clear, enough special cases arise to require lengthy guidelines, e.g., when is *The Wall Street Journal* an artifact, and when is it an organization? When is *White House* an organization, and when a location? Are branch offices of a bank an organization? Is a street name a location? Should *yesterday* and *last Tuesday* be labeled dates? Is *mid-morning* a time? For human annotator consistency, guidelines with numerous special cases have been defined for the Seventh Message Understanding Conference, MUC-7 (Chinchor, 1997).

In training data, the boundaries of an expression and its type must be marked via SGML. Various GUIs support manual preparation of training data and reference answers.

Though the problem is relatively easy in mixed case English prose, this is not solvable solely by recognizing capitalization in English. Though capitalization does indicate proper nouns in English, the type of the entity (person, organization, location, or none of those) must be identified. Many proper noun categories are not to be marked, e.g., nationalities, product names, and book titles.

Named entity recognition is a challenge where case does not signal proper nouns, e.g., in Chinese, Japanese, German or non-text modalities (e.g., speech). Since the task was generalized to other languages in the multi-lingual entity task (MET), the task definition is no longer dependent on the use of mixed case in English.

Broadcast news presents significant challenges, as illustrated in Table 1. Not having mixed case removes information useful to recognizing names in English. Automatically transcribed speech, even with no recognition errors, is harder due to the lack of punctuation, spelling numbers out as words, and upper case in SNOR (Speech Normalized Orthographic Representation) format.

Mixed Case The crash was the second of a 757 in less than two months. On Dec. 20, an American Airlines jet crashed in the mountains near Cali, Colombia, killing 160 of th 164 people on board. The cause of that crash is still under investigation.

UPPER CASE THE CRASH WAS THE SECOND OF A 757 IN LESS THAN TWO MONTHS. ON DEC. 20, AN AMERICAN AIRLINES JET CRASHED IN THE MOUNTAINS NEAR CALI, COLOMBIA, KILLING 160 OF TH 164 PEOPLE ON BOARD. THE CAUSE OF THAT CRASH IS STILL UNDER INVESTIGATION.

SNOR THE CRASH WAS THE SECOND OF A SEVEN FIFTY SEVEN IN LESS THAN TWO MONTHS ON DECEMBER TWENTY AN AMERICAN AIRLINES JET CRASHED IN THE MOUNTAINS NEAR CALI COLOMBIA KILLING ONE HUNDRED SIXTY OF THE ONE HUNDRED SIXTY FOUR PEOPLE ON BOARD THE CAUSE OF THAT CRASH IS STILL UNDER INVESTIGATION

Table 1: Illustration of difficulties presented by speech recognition output (SNOR).

3. OVERVIEW OF HMM IN IDENTIFINDER™

A full description of our HMM for named entity extraction appears in Bikel, et. al., 1997. By definition of the task, only a single label can be assigned to a word in context. Therefore, to every word, the HMM will assign either one of the desired classes (e.g., person, organization, etc.) or the label NOT-A-NAME (to represent "none of the desired classes"). We organize the states into regions, one region for each desired class plus one for NOT-A-NAME. See Figure 1. The HMM will have a model of each desired class and of the other text. The implementation is not confined to the seven classes of NE; in fact, it determines the set of classes by the SGML labels in the

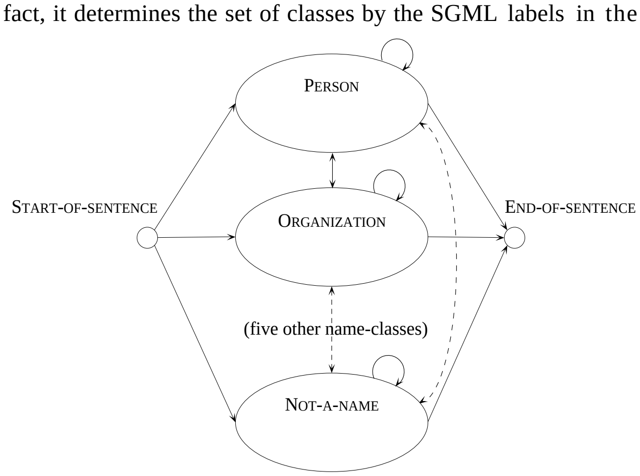

Figure 1: Pictorial representation of conceptual model.

training data. Additionally, there are two special states, the START-OF-SENTENCE and END-OF-SENTENCE states.

Within each of the regions, we use a statistical bigram language model, and emit exactly one word upon entering each state. Therefore, the number of states in each of the name-class regions is equal to the vocabulary size, $|V|$.

The generation of words and name-classes proceeds in the following steps:

1. Select a name-class *NC*, conditioning on the previous name-class and the previous word.
2. Generate the first word inside that name-class, conditioning on the current and previous name-classes.
3. Generate all subsequent words inside the current name-class, where each subsequent word is conditioned on its immediate predecessor.
4. If not at the end of a sentence, go to 1.

Using the Viterbi algorithm, we search the entire space of all possible name-class assignments, maximizing $\Pr(\boldsymbol{W}, \boldsymbol{NC})$.

This model allows each type of "name" to have its own language, with separate bigram probabilities for generating its words. This reflects our intuition that

- *There is generally predictive internal evidence regarding the class of a desired entity.* Consider the following evidence: organization names tend to be stereotypical for airlines, utilities, law firms, insurance companies, other corporations, and government organizations. Organizations tend to select names to suggest the purpose or type of the organization. For person names, first person names are stereotypical in many cultures; in Chinese, family names are stereotypical. In Chinese and Japanese, special characters are used to transliterate foreign names. Monetary amounts typically include a unit term, e.g., Taiwan dollars, yen, German marks, etc.

- *Local evidence often suggests the boundaries and class of one of the desired expressions.* Titles signal beginnings of person names. Closed class words, such as determiners, pronouns, and prepositions often signal a boundary. Corporate designators (Inc, Ltd., Corp., etc.) often end a corporation name.

While the number of word-states within each name-class is equal to $|V|$, this "interior" bigram language model is ergodic, *i.e.*, there is a probability associated with every one of the $|V|^2$ transitions. As a parameterized, trained model, if such a transition were never observed, the model "backs off" to a less-powerful model.

4. EVALUATION MEASURES

Information extraction from text is measured in terms of precision (P) and recall (R), terms borrowed from the information retrieval community, where

$$P = \frac{\text{number of correct responses}}{\text{number of hypothesized responses}} \qquad \text{and}$$

$$R = \frac{\text{number of correct responses}}{\text{number of tags in reference}}.$$

The F-measure is the uniformly weighted harmonic mean of precision and recall:

$$F = \frac{RP}{(R+P)/2}$$

In MUC and MET, a correct response is one where the label and both boundaries are correct. A response is half correct if the type is correct and the response string overlaps with the reference string. Alternatively, a response is half correct if the class of the type (rather than the type) and both boundaries are correct. Type classes are defined as follows:

entity (ENAMEX): PERSON, ORGANIZATION, LOCATION

time expression (TIMEX): DATE, TIME

numeric expression (NUMEX): MONEY, PERCENT.

Scoring NE on speech is not merely a matter of applying the MUC scoring algorithm, since it assumes that the source text of answer key and system output are identical. One needs to allow for insertion/deletion/substitution errors by a speech recognizer and compare that against one reference answer.

In 1997 we developed a procedure for scoring NE on speech. As shown in Figure 2, first one aligns the speech recognizer output (HYP) to the reference text (REF), then merges the NE annotation from the system to the aligned HYP, and merges the NE answer key to the aligned REF. We used the word alignment software from NIST. The aligned NE-annotated HYP and REF can be scored using the MUC scorer.

An alternative (Burger, Palmer, and Hirschman, 1998) from MITRE is under development. It has several virtues. First, it is flexible, allowing alternative alignment strategies, not just word alignment. Second, it is more forgiving of speech errors. It computes a separate measure for content variance, e.g., "SMYTHE" in the reference answer (REF) versus "SMITH" in the hypothesis (HYP). It allows one to specify weights on the different kinds of errors; one can replicate the current MUC metrics with appropriate weights.

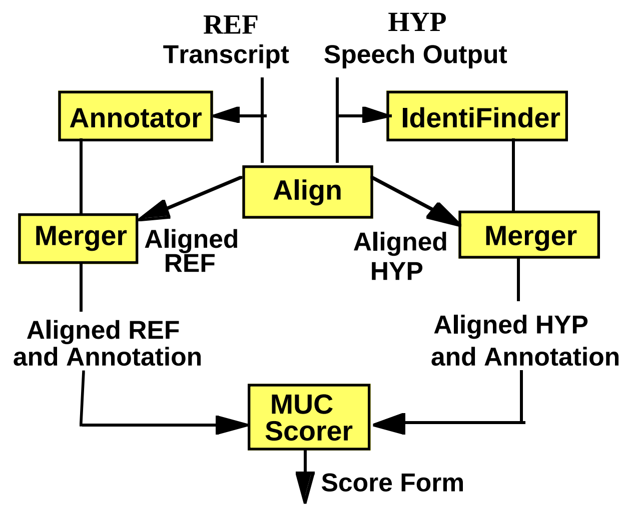

Figure 2: Evaluation using NIST alignment software and official MUC scoring software.

5. RESULTS AND ANALYSIS

All results reported are based on the MUC scoring metrics and software (Chinchor, 1997) so that NE performance on speech can be compared to results published on text.

5.1 Annotation

Annotation involved two individuals independently marking LDC data, which is in SNOR format. Those two were scored against each other, typically yielding 97% agreement among the annotators. All differences were adjudicated by a third person. To further look for inconsistencies, we then trained IdentiFinder on the new data and deliberately tested on training. Scoring IdentiFinder against the adjudicated keys typically resulted in scores of 99. Any discrepancies were reviewed by the adjudicator, resulting in a gold standard.

We annotated 467k words of training. Including double annotation, adjudication, test-on-training, readjudication, and management, this took roughly 26 person days. The annotators were college students; the adjudicator happened to be a graduate student.

We then ran an experiment to see whether having two independent annotators was important. It demonstrated that if we had used only one highly practiced annotator, F-scores would have been affected only by 0.4. Therefore, one highly practiced annotator, followed by test-on-training, followed by adjudication would have been just as good as our process, but would have required only 12 person days by our estimates.

5.2 Effect of Training Set Size

Previously, we had observed that performance on text input was remarkably good with as little as 100k words of training data (Bikel, et. al., 1997), though performance continued improving with additional training.

To assess the effect of training set size with broadcast news, we trained IdentiFinder on three successively larger training sets, testing each time on held out SNOR transcripts from the Linguistic Data Consortium (LDC). The results of the three runs appear in Table 2.

Performance with as little as 82,000 words (only 5,000 tags of NE data) is surprisingly good. Performance increases as training increases, though the gain from doubling training seems to decrease exponentially.

No. of Words	No. of Tags	F-Measure
82k	5k	82
212k	12k	86
467k	25k	87

Table 2: Effect of increased training on performance.

It is worth noting that the density of NE exemplars in broadcast news is only half that in newswire. Therefore, though the training for text that should not be marked is adequate, the number of examples for named entities is only about half what we would like to see.

To look at the impact of training another way, we counted the errors based on the frequency of the exact NE string appearing in training; the training consisted of 467,000 words containing 25k tags of NE strings. See Table 3.

No. of examples	Errors / total	Error rate
0	313/710	44%
1-63	38/817	5%
63-1023	7/520	1%

Table 3: Error rate as a function of frequency of occurrence in training.

As expected, the error rate on NE strings never seen in training is highest (44%). Furthermore, the error rate on previously unseen NE strings is 87% of our errors.

Perhaps surprising, even seeing the NE string in training once dramatically reduces error; the error rate is a fairly constant 5% for strings observed from 1-63 times in training.

This suggests two points:

- There is an opportunity to improve overall by improving performance on sequences not seen in training.
- The reduction in error rate from additional training is largely coming from reducing the fraction of sequences never seen in training.

5.3 Performance on Broadcast News versus Newswire

In this section we compare performance in the broadcast news domain versus newswire. Our comparison is based on the New York Times News Service (NYT), which is being used in the Seventh Message Understanding Conference (MUC-7); the Wall Street Journal (WSJ), which was used in MUC-6; and on the HUB-4 data distributed by LDC in 1996 and 1997.

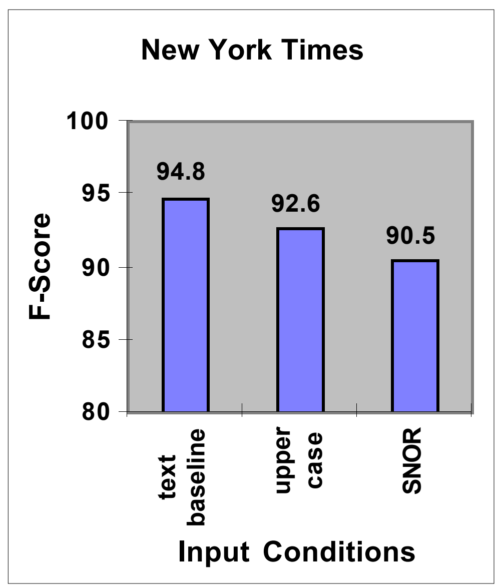

Figure 3: Performance variation on NYT, given mixed case text, versus upper case, versus SNOR formatted input.

Figures 3 and 4 measure the effect of not having case and of SNOR input in the New York Times newswire (NYT) and in the Wall Street Journal (WSJ) respectively. The baseline us well-written mixed case prose. Using the same test materials, but in uppercase, we see that performance degrades somewhat on the upper case version. The degradation is less in WSJ than in NYT in part we believe, due to style conventions in WSJ. For example, the first mention of a person is usually the full name, e.g., *John Doe;* subsequent mentions of the individual tend to include a title, e.g., *Mr. Doe.*

Using the same test materials but in SNOR format, after training on SNOR text, performance from upper case to SNOR is about two points of F-score in both NYT and WSJ.

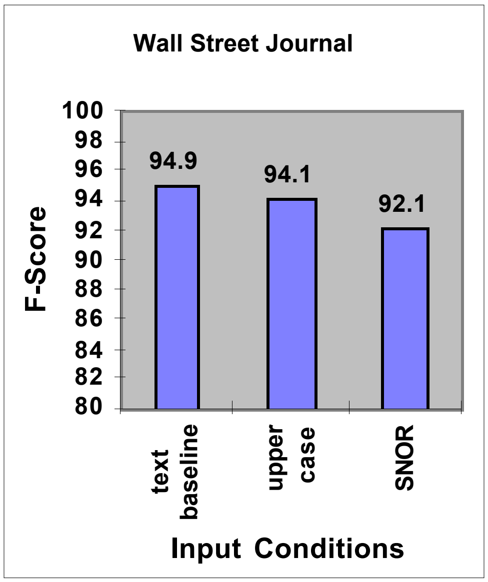

Figure 4: Performance variation on WSJ, given mixed case text, versus upper case, versus SNOR formatted input.

Two differences have already been noted:

- In broadcast news, NE exemplars are only half as dense as in newswire.
- In SNOR format, there is no case, punctuation, nor numbers written with digits.

There are other obvious differences. NYT is consistently formatted and well written. Broadcast news, on the other hand, contains carefully composed sections read by a news anchor, spontaneous speech, interviews, and dialogs. As a result, there are less formal sections that include disfluencies.

Table 4 contrasts performance on NYT test data with performance on broadcast news in SNOR format. In each case, IdentiFinder had been trained only on the appropriate source, e.g., NYT training for the NYT test and broadcast news training for the broadcast news test.

The factors cited regarding the differences between NYT and broadcast news suggest that performance on broadcast news should be much worse than on newswire. However, it is pleasantly surprising how little performance degrades.

Since we have only SNOR documents, we cannot be sure how much of the degradation is due to content versus how much is due to SNOR format. Roughly, four points of F seem lost due to SNOR (see Figure 3). We speculate that as much as 3 points of F are due to the broad domain of broadcast news (the NYT test for MUC-7 is on a specific domain, air disasters, rather than all news) and due to the segments of spontaneous speech, rather than carefully edited prose.

Source	Text Type	F	Comments
New York Times	Normal	95	Consistently formatted
New York Times	SNOR	91	No case; no punctuation
Broadcast News	SNOR	87	Spoken domain; less formal; SNOR; manual transcription

Table 4: Effect of broadcast news domain and SNOR contrasted with newswire.

5.4 Effect of Recognition Errors

In Table 5, we show the effect of speech recognition errors. In a preliminary experiment in the spring of 1997, using the '96 HUB-4 development test, we noticed a drop in performance from an F of 82 to 60 given a word error rate of 30% in the speech recognizer. That degradation is less than one might expect. For instance, the average length of names in that data is 1.7 words. However, the degradation in F-score is less than the word error rate would predict on single words (e.g., 82 x 0.7 = 57.4). Performance (in terms of F-score) is sensitive to speech recognition performance and linear with respect to anticipated improvements in word error rate. An analysis of the errors made with speech recognition input showed that the dominant error was missing names; the second most prominent error is spurious names.

Source	WER	F
'97 SNOR REF	0%	87
'97 HYP	20%	73
'96 SNOR REF	0%	82
'96 HYP	30%	60

Table 5: Degradation in performance due to speech recognition errors.

We reran the experiments recently using the '97 HUB-4 data. Since IdentiFinder improved during the last year, the baseline on the '97 SNOR REF data is 87, contrasted with the lower score a year ago. However, the degradation from speech recognition errors again was less than the 20% word error rate would predict (87 x .8 = 70). The average length of name strings was 1.8. Most new errors that occurred with speech input, rather than the error-free transcripts, were due to missing names during recognition.

In both cases, NE performance degraded roughly proportional to the speech recognition errors.

6. RELATED WORK

A variant of automatically learning Brill rules has been applied to the NE problem, as outlined in (Aberdeen et al., 1995). Performance thus far reported is in the mid 80s on the MUC-6 (The Wall Street Journal) data contrasted with IdentiFinder's performance in the mid-90s. More recently, Burger et. al., 1998, report results of applying Brill rules to speech.

Bennett (1997) reports on using binary decision trees a la C4.5 for the NE task. The decision tree decides whether to insert a begin category mark, end category mark, or nothing at each point in the sequence of input words. Their scores thus far are about 91, contrasted with IdentiFinder's 94.5 on the same test material. The technique has thus far been applied only to newswire text.

These two alternative learning approaches use statistics to make decisions. Only IdentiFinder has a complete probabilistic model that

- governs all decisions and
- models the categories of interest and the residual input that is not of interest.

7. CONCLUSIONS

We have drawn the following conclusions:

1. Broadcast news represents a more difficult domain than newswire.
2. Degradation in performance from mixed case to SNOR is relatively small.
3. Degradation from speech recognition errors seems to directly track word error rate.
4. Evaluations of speech based on named entity extraction appear to be interesting and complementary to word error rate.

ACKNOWLEDGMENTS

The work reported here was supported in part by the Defense Advanced Research Projects Agency. Technical agents for part of this work were Fort Huachucha and NRaD under contract number DABT63-94-C-0062 and N66001-97-D-8501. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the United States Government.

We appreciate the assistance of Lynnette Hirschman, David Palmer, and others at MITRE for their scoring software; Nancy Chinchor and Aaron Douthat of SAIC for the MUC scoring software; and Ann Albrecht, Dan Bikel, Michael Crystal, Georgina Garcia, and Scott Miller of BBN.

REFERENCES

1. Aberdeen, J., Burger, J., Day, D., Hirschman, L., Robinson, P. and Vilain, M. (1995) "MITRE: Description of the Alembic System Used for MUC-6". In *Proceedings of the Sixth Message Understanding Conference (MUC-6)* Morgan Kaufmann Publishers, Inc., Columbia, Maryland, pp. 141–155.
2. Bennett, S.W., Aone, C., and Lovell, C. (1997) "Learning to Tag Multilingual Texts Through Observation." In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, Providence, Rhode Island, pp. 109-116.
3. Bikel, D., Miller, S., Schwartz, R., and Weischedel, R. (1997) "NYMBLE: A High-Performance Learning Name-finder". In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Association for Computational Linguistics, pp. 194-201.
4. Burger, J.D., Palmer, D., and Hirschman, L. (1998) *Named Entity Scoring for Speech Input*, to appear.
5. Chinchor, Nancy (1997) *MUC-7 Named Entity Task Definition Dry Run Version, Version 3.5, 17 September 1997* Available via ftp or telnet at *online.muc.saic.com*, in the file *NE/training/guidelines/NE.task.def.3.5.ps*.